contamadrid - Predicción de niveles de NO2 en Madrid

Luz Frias

August 15, 2016

Contents

Propuesta de valor

Durante el último año, el tema de la contaminación en Madrid ha adquirido gran protagonismo. En parte por la creciente preocupación por el medio ambiente y el efecto de la contaminación en la salud pública, y en parte por las recientes restricciones de circulación impuestas por el Ayuntamiento de Madrid para disminuir los niveles más altos.

Objetivo del proyecto

Hasta ahora, estas restricciones se aplican de manera reactiva, es decir, tras haber alcanzado niveles altos de contaminación. Este proyecto está motivado por intentar abordar el problema de forma proactiva, prediciendo con al menos un día de antelación niveles de alerta de NO_2.

Portal web

Los resultados son:

- El portal web de contamadrid.es
- El desarrollo de una API abierta para que otros desarrolladores integren soluciones consultando los resultados calculados en este proyecto. Por ejemplo, para la creación de aplicaciones móviles (Android, iOS, ...)

Metodología

En la construcción de todos los modelos predictivos de este proyecto, se realiza uno por estación metereológica, ya que sus características innatas influyen enormemente en la evolución de la contaminación. Por ejemplo, el tráfico medio influye significativamente más en las estaciones urbanas (como Plaza de España) que en las instaladas en grandes zonas ajardinadas (como Casa de Campo). Además, esto hace que el proyecto sea fácilmente generalizable, permitiendo la fácil incorporación de datos de otros municipios.

Datos

Contaminación

Los datos de contaminación los publica el ayuntamiento en su portal de datos abiertos. Los datos se pueden encontrar con dos periodicidades diferentes (horaria y diaria) y en formato histórico o tiempo real (datos del día actual, con actualización horaria y un retraso de aproximadamente una hora).

En este proyecto me he centrado en el estudio de los datos a nivel horario. Algo curioso es que no hay problemas en encontrar el histórico desde 2001 hasta el mes previo al actual, ni los datos de hoy, pero no se publican en ningún lado los datos del mes actual hasta el día previo a la consulta.

El parseo de los ficheros resulta un poco incómodo, ya que, aunque contienen la misma información el histórico y los datos en tiempo real, tienen formatos diferentes:

- El histórico se guarda en un fichero de anchos fijos
- El de datos en tiempo real, en un fichero separado por comas

Además que en cada fila se representa un día, y los datos horarios están en formato ancho (un valor horario en cada columna). Por conveniencia para el entrenamiento, el proceso de limpieza lo pasa a formato largo (una fila por hora).

Mediciones históricas

Se encuentran aquí zipeados por año. Para facilitar reprocesos (durante el análisis detecté que faltaban algunos meses, y la mayoría de ellos han sido repuestos tras avisar al administrador del portal) y la inclusión de nuevos datos, los datos se recogen mediante un proceso que:

- Descarga los zip
- Los descomprime
- Los lee y pasa los procesos de limpieza necesarios
- Los guarda como un fichero por año

Mediciones en tiempo real

Se encuentran aquí. Se ha desarrollado un proceso que:

- Lo lee y pasa los procesos de limpieza necesarios
- Lo guarda como un fichero por día

Información de estaciones

Se encuentra aquí en formato excel. Lo he transformado en un proceso semi-manual a texto plano, transformando las coordenadas de formato grados, minutos y segundos a formato decimal.

Variables predictoras

Introducción

El primer paso fue estudiar cuáles son las variables que influyen en la varación de los niveles del NO_2. Tras algo de documentación en páginas relacionadas y papers científicos, se determina que son:

- Las fuentes principales son las emisiones de vehículos e industriales
- La variación del nivel está relacionada con factores metereológicos, especialmente:
  - Velocidad del viento
  - Humedad relativa
  - Temperatura

Metereología

Los datos utilizados para este proyecto se han obtenido de OGIMET, que proporciona registros diarios y horarios para un conjunto reducido de estaciones. He escogido la estación de Madrid - Barajas por ser la más cercana a la ciudad de entre las disponibles.

Los datos incluyen:

- A nivel diario:
  - Temperatura media, mínima y máxima
  - Humedad relativa
  - Velocidad del viento
  - Nivel de precipitaciones
- A nivel horario:
  - Temperatura

Niveles de tráfico

También disponible en el portal de datos abiertos del ayuntamiento de Madrid (aquí) se puede encontrar un histórico de los niveles de tráfico en la M-30 desde 2013, que se actualiza diariamente.

Vienen datos totales en formato xml con un nodo por día.

La inclusión de los datos de niveles de tráfico, supone en general una mejora de la predicción, pero "perdemos" el histórico de 2001 a 2013 que sí tenemos disponible en datos de contaminación y metereología.

Predicción

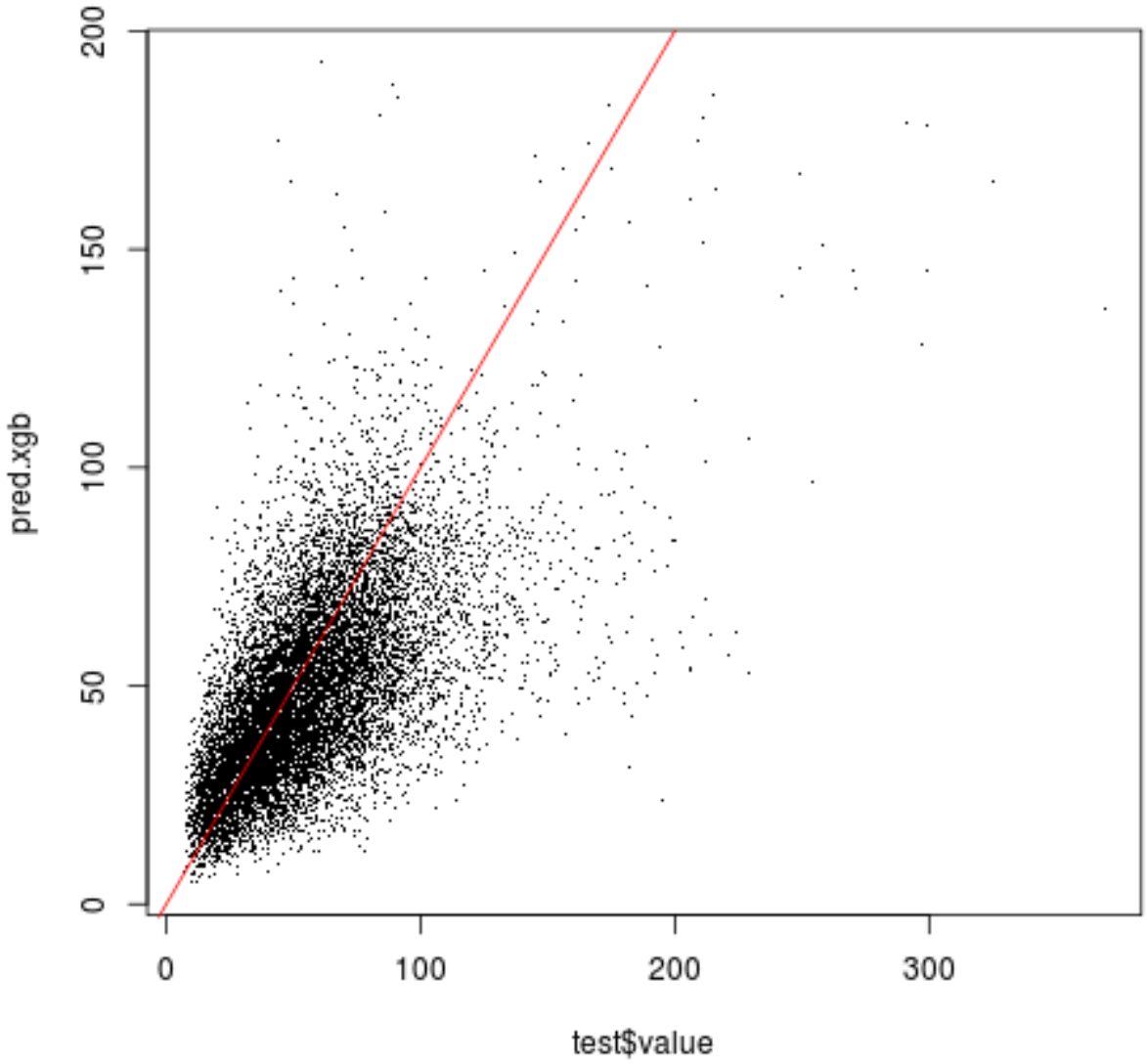
Debido a las características de cada punto de medición, en todos los casos se ha creado un modelo por cada estación. Un ejemplo es cómo influye el nivel de tráfico a estaciones cercanas a puntos de alta densidad de circulación con respecto a puntos situados en grandes zonas ajardinadas (p.ej. parque Rey Juan Carlos I).

Valores de NO2 con Gradient Boosting Trees

Los modelos realizados tienen las siguientes características:

- Paquete de R xgboost.
- Entrenados mediante validación cruzada.
- Los datos incluidos en cada iteración del entrenamiento durante la validación cruzada es un subconjunto pequeño de los datos (7.5%). Esto es para evitar overfitting, ya que los datos de un mismo día pero diferentes horas tienen muchos datos en común (intensidad tráfico, datos metereológicos a nivel diario, ...) y con árboles profundos tiende a memorizar los valores fijando los datos comunes.
- Se sobrescribe la métrica de evaluación a minimizar (por defecto RMSE) por una customizada, que penaliza los errores en valores altos. Es decir, es peor equivocarse por 30 microgramos en un valor real de 180 que en uno de 20. Aunque haciendo la prueba de volver a entrenar optimizando el RMSE los resultados son prácticamente los mismos.
- Tuning de parámetros mediante caret.
- Separación de conjunto de entrenamiento y validación por fecha, intentando simular que predecimos valores futuros en base a observaciones pasadas.

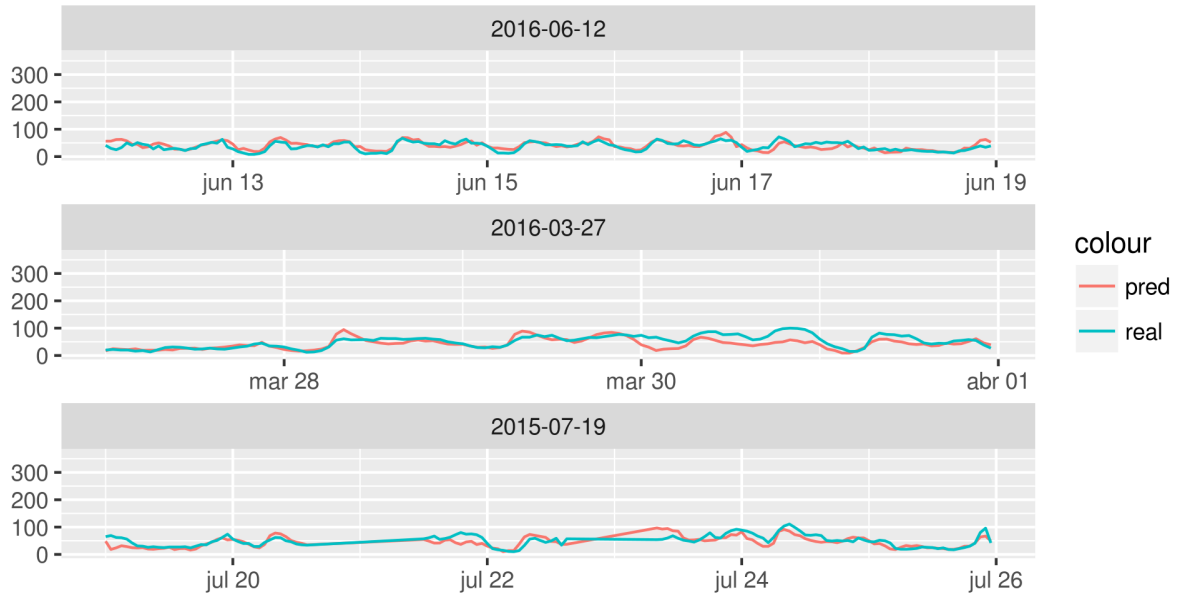
Los resultados son en general buenos para valores normales, pero se pierden a veces los picos. P.ej. esta es la comparación entre valores reales y predichos en la estación de C/Alcalá con C/O'Donell.



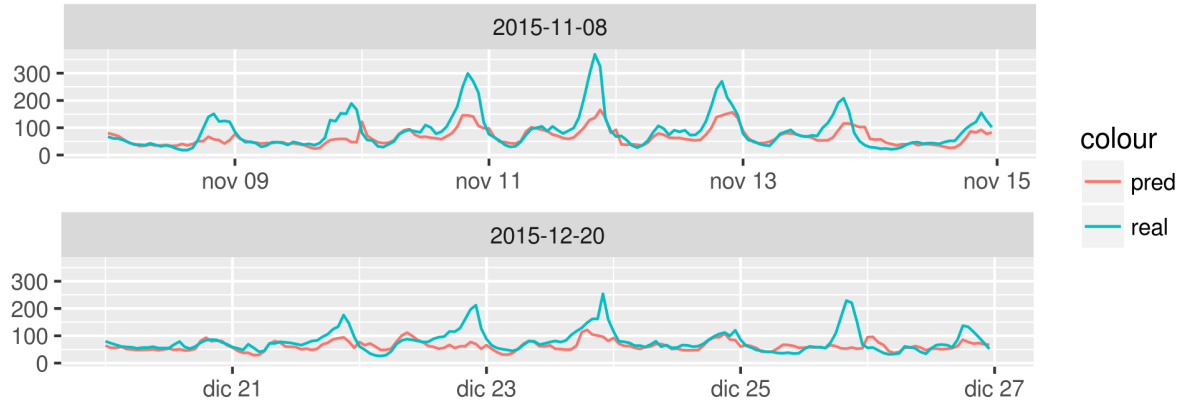
El rendimiento general sobre los datos de validación es:

- MAE: entre 9 - 19
- RMSE: entre 12 - 27

Aquí se puede ver cómo los valores bajos y normales se predicen muy bien:



Y aquí otras no tan precisas:



Clasificación de nivel de aviso con Random Forest

Lo siguiente es una simplificación del problema. En lugar de intentar predecir el valor horario del NO_2, transformamos la pregunta a si en un determinado día se va a superar o no el nivel de preaviso (180 microgramos/m³ durante 2 horas) y de aviso (200 microgramos/m³ durante 2 horas).

Los modelos realizados tienen las siguientes características:

- Paquete de R randomForest.
- Separación de conjunto de entrenamiento y validación por fecha, intentando simular que predecimos valores futuros en base a observaciones pasadas.
- Al contrario que en los modelos anteriores, no usamos el nivel de tráfico, y así podemos entrenar con todo el conjunto de datos.
- Se agrupan los datos a nivel diario.
- Se incluye como variable predictora el percentil 95 de los niveles del día anterior (para tener en cuenta el efecto acumulativo).

- Como el modelo está muy desbalanceado (muchas observaciones negativas por cada una positiva), entrenamos 25 modelos, en el que cada uno incluya dos partes de observaciones negativas por cada una positiva, y se combinan. Se hace 2 - 1 en lugar de 1 - 1 porque se ha observado una mejora en los resultados de esta forma. Tiene sentido, que en caso de duda se "vote" al no.
- Se ha determinado el punto de corte de la probabilidad a partir de la cual se considera el sí en la predicción de la siguiente forma:
  - Se predice utilizando los valores del 0.50 al 0.95 con incrementos de 0.05
  - Se calcula el coste del error, penalizando el falso negativo 4 veces más que el falso positivo.
  - Se escoge la probabilidad con menor coste

Tecnología empleada

Software

El proyecto consta de las diferentes piezas software, y cada una de ellas hace uso de distintas tecnologías:

- Procesamiento y analítica de datos: R.
- Back-end web (API): Python con Bottle.
- Front-end web: HTML, CSS, backbone y leaflet.
- Control de versiones: git

Hardware

Todas las piezas del proyecto están desplegadas en máquinas de Amazon Web Services, concretamente en instancias de EC2 con Ubuntu Server.

Viabilidad del proyecto

El proyecto ya ha sido desarrollado con resultados satisfactorios tanto para:

- Predecir la evolución horaria del NO_2
- Detectar días de riesgo, con probabilidad significativa de alcanzar niveles altos de este contaminante

Expectativas

El proyecto no termina aquí, los próximos pasos a acometer a medio plazo son:

- Combinación de modelos para mejorar la predicción
- Extender el análisis a otros contaminantes como el O_3, CO y las PM
- Desarrollar una app con los resultados generados y disponibles en la API de contamadrid. Esta app incluirá notificaciones push configurables por el usuario, de modo que le alerte cuando el valor real o predicho supere cierto nivel